会话分析视角下的突发公共事件主题演化研究*
——以“新冠肺炎疫情”为例

翟姗姗 王左戎 陈欢 潘港辉
华中师范大学信息管理学院 武汉 430079

摘 要:[目的/意义]会话分析理论的引入为主题演化研究提供了新的研究视角,细化了主题演化分析粒度。同时,更为完善的主题演化分析思路被应用于突发公共事件之中,有利于提升监管部门的舆情疏导效率。[方法/过程]针对现有研究中的主题识别方法与主题演化判断标准,结合会话分析与主题分析,将会话内容与会话组织结构引入主题演化分析过程中,并以“新冠肺炎疫情”中用户生成内容(UGC)作为数据来源进行实证分析。通过基于时序性与讨论热度的主题演化分析,从主题强度层面识别不同层级内容的演化规律,并在主题内容分析层面引入知识发现的关联规则计算思想以挖掘语料内容间的参照关系,结合社会网络分析方法确定关键演化路径。[结果/结论]研究结果表明,网络结构中不同层级的主题内容存在一定差异并对主题演化趋势有着重要影响,对有着重要作用的层级的内容进行有效监管会对引导舆情走向产生积极作用。

关键词:会话分析 突发公共事件 主题识别 主题演化 关联规则
分类号:G206
DOI:10.13266/j.issn.0252-3116.2022.11.010

1 引言

近年来,诸如“新冠疫情”“7·20郑州特大暴雨”等突发公共事件对社会稳定与经济发展均产生了重大影响。随着网络技术的发展和移动智能设备的普及,微博、微信、短视频、在线社区等非正式信息交流方式受到了广泛青睐,公众在线参与事件舆情讨论的意愿日益强烈。与传统媒体舆论传播相比,大数据时代下的网络舆情具有信息相对开放、信息传播迅速、信息丰富多样和信息具有倾向性的特点[1]。与规范化的新闻语料或政策文本不同,由用户自己生产、更新并借助于网络媒介传播的大量用户生成内容(User Generated Content,UGC),不仅能全面刻画用户更深层次的内容偏好,且其中所蕴含的主题及主题演变趋势与演化规律,有助于准确了解与掌握突发公共事件的脉络现状、发展规律与动态趋势,为相关部门实施智能监控、辅助决策、舆情引导、个性化推荐等提供了服务参考。当前,主题演化方面的研究已得到多方面的深化与拓展,但仍受到分析层次宏观与测度指标维度单一等方面的局限。而会话分析作为一种有效揭示非正式信息交流内容、关系、行为等规律的有效手段[2],不仅为从信息交流数据中探究反映人类言语交互的社会学规律提供了理论依据,更为以UGC数据为来源、以主题为表征、以识别主题内容潜在关联关系为目的的主题持续演化规律研究提供了特定场景。

基于此,本文以突发公共事件为研究对象,结合会话分析与主题分析,将会话内容与会话组织结构引入主题演化分析过程中,将“新冠肺炎疫情事件”UGC数据视为基于社交媒体的异步会话过程开展实证分析。一方面,通过基于时序性和讨论热度的主题演化分析,借助于会话内容呈现的层级结构探寻不同层级主题的演化规律;另一方面,利用基于关联规则的支持度和可信度,通过判定主题内容间的语义关联和演化趋势,从而确定主题关键演化路径,为后续舆情引导、监控、管理与预测提供参考性建议。

2 相关研究

2.1 突发公共事件主题识别与演化的相关研究

与一般公共事件所不同,突发公共事件具有突发性和高破坏性的典型特征,对其进行主题识别与主题

* 本文系国家社会科学基金青年项目“大数据视角下突发公共卫生事件信息协同体系研究”(项目编号:20CTQ021)研究成果之一。
作者简介:翟姗姗,副教授,博士,E-mail:zhais@mail.ccnu.edu.cn;王左戎,硕士研究生;陈欢,硕士研究生;潘港辉,硕士研究生。
收稿日期:2021-10-29 修回日期:2022-01-17 本文起止页码:87-99 本文责任编辑:王传清

演化分析不仅是突发公共事件舆情领域中的重要研究内容,更对未来的舆情监管工作具有一定的借鉴意义。针对突发公共事件主题识别方法的研究,主要涉及共词分析和概率模型两个方向。第一类,以共词分析为基础的主题识别方法,是利用文本集合中词与词之间的共现关系来反映关系强度,从而实现主题聚类与识别的。其具体应用场景包括挖掘潜在主题[3]、解决自标引关键词缺失[4]等。第二类,以概率模型为基础的主题识别方法,其实现原理的核心是机器学习算法,例如早期的 LDA(Latent Dirichlet Allocation)模型。其后的研究一方面围绕基本 LDA 主题模型进行,包括新闻文本[5]、UGC 数据[6]在内不同类型数据源的针对性拓展,另一方面从主题识别的各个环节入手,不断提出基于 LDA 的优化模型,如 SECNN 模型[7]。

突发公共事件主题演化的相关研究主要侧重于对提取出的主题进行语义层面的相似度计算来推进演化分析。在主题内容方面,有学者借助社会网络分析工具构建事件语义图谱,进一步提出基于语义的突发公共卫生事件网络舆情主题发现框架并进行实证分析[8]。还有学者将突发公共事件中利益相关者的话题关注点进行分类分阶段分析,揭示不同利益相关者在不同阶段话题演化模式的异同点[9];主题强度方面,有些学者针对公共政策数据进行采集分析,通过结合 LDA 主题模型与离散时间法,借助讨论热度等多项指标比较不同类型主题政策的演化情况[10]。部分学者以舆情传播四阶段为基础进行划分,分别针对各个阶段进行主题的提取与演化分析,并提出了基于微博文本的舆情管控策略[11]。除了从单一层面出发的研究以外,还有学者针对从知乎平台中获取到的数据来提取主题,总结不同时段内用户关注的重点主题内容,并归纳不同主题内容的主题强度变化趋势[12]。部分学者针对新冠肺炎疫情期间的谣言这一特殊对象进行主题内容分布特征以及数量特征的剖析,并结合马斯洛需求理论分析其深层次的形成原因[13]。还有学者选择从话题讨论数量、热度及内容等多维特征出发全方位追踪舆情话题的演化情况,并基于知识图谱方法构建话题图谱[14]。随着社交媒体技术的发展,舆情内容成为了人民真实意愿的直接表现。舆情监测过程中对情感倾向的识别也成为了政府相关部门以及研究者的关注焦点。主题情感相关的研究大多是基于主题内容或主题强度的分析,结合包括 VADER 情感模型[15]、情感单元词表[16]在内的相关情感词典来进行的。

综合现有的相关研究文献可见,在主题识别方面,现有方法较少考虑语料内容在语义层面的关联关系;在主题演化分析方面,或大多聚焦于主题的时序变化与热度演化趋势,或直接通过主题间的文本相似度来度量主题关联程度,仍缺少对于主题间关联关系的有向性判定标准的关注。

2.2 会话分析及应用研究

会话分析(Conversation Analysis),即对日常生活中自然发生的真实会话进行记录和分析,该理论认为人们日常会话的构成是存在一定秩序和规律的。会话分析视角下的研究,一方面聚焦于人们会话过程中的语言表达,并通过分析具体的语用和语义特征,总结话语角色的会话风格和会话策略,并应用于诸如课堂互动[17]、医患交流[18]、心理咨询[19]、综艺节目赏析[20]和市场交易[21]等多个场景中。

另一方面的研究则更为关注会话语料集合来源所带来的话语角色间的组织结构和交互模式异同。随着网络技术与新媒体的发展,更多的会话语料从线下转向线上,以在线社区为代表的非正式交流成为其主要语料来源。例如,李纲等综合社会网络分析法和内容分析法对所收集的微信群内的语料进行剖析,并构建出群聊内参与会话成员的交流网络[22-23]。部分学者选择线上学术社区作为研究对象,结合 LDA 主题模型对信息交互类型及内容拓扑结构等进行深入分析,并提出针对性的促进虚拟学术社区用户交互的策略[24]。李月琳等则将研究目标转变为因受到疫情影响而受到关注的医疗健康网站,通过医疗健康网站中医生和患者的会话轮次等交互数据分析影响交互效率的因素,为后续相关平台或系统的开发提供理论指导[25]。

综上,现有的突发公共事件主题演化分析方面的研究,或基于时序关系、或基于主题热度或基于主题相似度度量主题之间的相关性,相关研究存在以下几个方面的局限:①分析层次较为宏观,一般将 UGC 视为一个大规模数据集,直接从中提取主题,忽略了由社交媒体或在线社区本身存在的网络组织结构所带来的影响,如未考虑“主帖 - 回复帖 - 楼中楼帖”中存在的主题内容差异;②主题内容关联方法较为单一,普遍采用语义相似度或语义距离的度量方式,既缺乏对于语料内容关联关系语义层面的理解,也未考量主题内容关联的有向性;③测度指标选择上理论依据不足,或基于时序、或基于主题热度、或基于主题文本相似度计算,尚未针对于主题持续性演化规律制定多维度测度标准。而会话分析的引入,既通过会话内容呈现了 UGC 数据中的层级结构关系,有利于探究不同层级主题的

演化规律，也实现了对于语料内容分析处理的细粒度化，将主体关联深度从内容层面延伸至语义层面，丰富了主题演化持续性判定的标准。因此，本文在会话分析视角下充分考虑语料内容的组织结构，在主题强度层面挖掘不同层级主题间的演化规律；同时在主题内容层面引入知识发现中关联关系的计算思路，在计算关联强度的同时实现对于演化关系指向性的判定与划分，同时结合社会网络分析方法识别关键演化路径，以揭示事件的主题演化规律。

3 研究方案设计

3.1 整体研究思路

社区类交流应用程序的出现改变了数据的单向流动现状并进一步丰富了活动产生的数据内容。本文最终选择在线交流社区百度贴吧中的“新型冠状病毒”吧作为原始数据源，其原因具体包括以下 3 个方面：①从信息发布者的角度而言，作为拥有超过 10 亿注册用户的百度贴吧，从中获取的 UGC 可以充分反映广大群众的内心真实需求；②从信息产生的渠道而言，UGC 数据相比较于仅来源于新闻或官方机构网站发布的权威类数据而言，数据来源更为广泛；③从信息呈现的形式而言，本文认为 UGC 数据所具有的“主帖－回复帖－楼中楼帖”层级结构是主题演化的必要因素，也是当前在线交流社区用户间交互所呈现出的主要模式。

本研究首先根据数据源特点建立爬虫框架进行数据爬取，并进行基础的预处理操作，包括空值处理、去除停用词、分词处理等。其次利用主题模型对每个帖子进行主题提取并进行聚类形成主题簇，根据特征词与主题的对应关系将帖子分配至对应主题。在主题演化分析阶段，一方面通过统计主题簇出现频次总结主题簇在时间维度上的热度变化，并结合具体共现关系研究不同主题簇之间的交互关系；另一方面聚焦特征词对的关联关系，并映射至主题簇层面来挖掘主题簇语义层面的关键演化路径。两者均是主题演化的重要反映，前者聚焦于主题被关注程度的趋势变化特征，后者则关注在突发公共事件推进过程中主题在内容层面的深入或延伸，具体研究思路如图 1 所示：

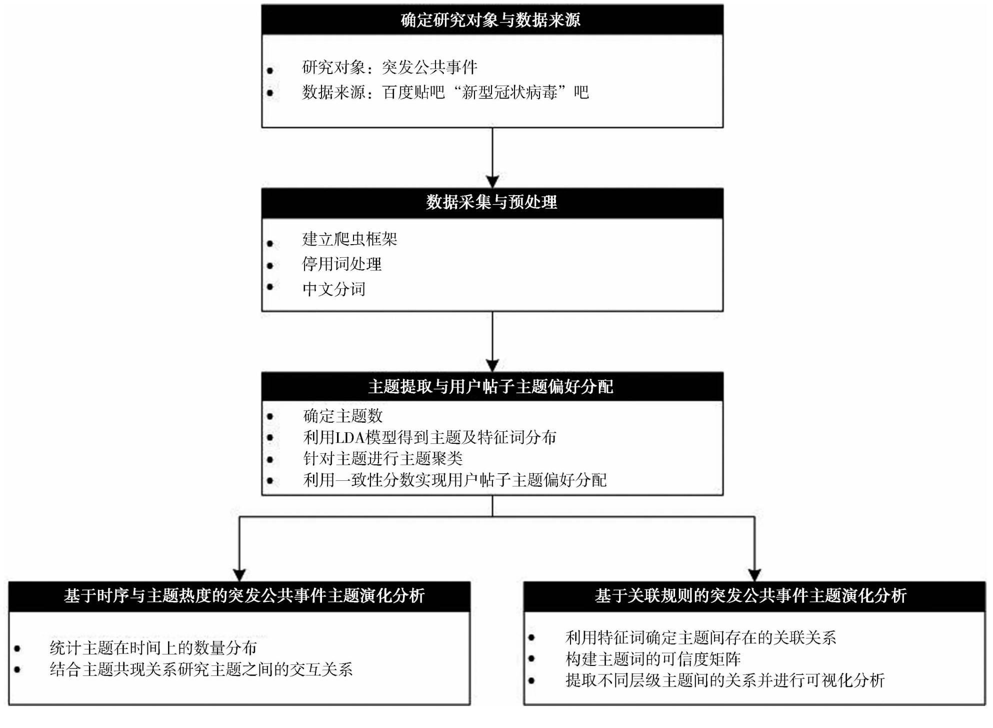


图 1 研究思路

3.2 研究方法

3.2.1 主题模型

主题模型是一种对文本隐含主题进行建模的方法，通过将高维度的词的集合映射到低维度的主题空间上来实现对目标数据的降维，建立简洁的表示[26]。在已有的研究当中，主题模型可以根据适用对象的不同分为两种。第一种是面向长文本的主题模型。典型代表有 LDA 主题模型以及后续对此进行优化的动态主题模型[27]和 TOT 模型[28]。但上述模型大多适用于长文本的处理，对于内容较少的短文本来说会出现数

据稀疏的问题。因此,第二类主题模型主要面向的是不超过 10 个词的短文本。例如 BTM 模型[29]和词网络主题模型 WNTM[30]。

考虑到研究过程中使用的帖子数量以及单个帖子的篇幅,本文最终选择 LDA 主题模型用于主题提取。首先,LDA 模型的主要特点之一是其能够从海量的文本信息中提炼出有效信息,并且能够为每一条信息分配对应的主题。其次,LDA 主题模型运行过程中所涉及的先验概率分布可以有效规避机器学习过程中产生的过拟合问题。最后,LDA 模型运行结果中还包含特征词隶属于主题的概率,可以满足研究中后续步骤的需要。另外,尽管现有研究中提出优化后的 LDA 模型在提取准度和算法性能上兼具优势,但在应用场景或语料对象上仍存在局限性,并不能够完全适配本文的研究场景。

3.2.2 主题聚类方法

主题聚类,即对某一特定特征差异较小的主题进行归并,并形成相对应的主题簇。为了使得主题提取的结果更为精确,本文需要根据主题模型提取结果进行聚类分析,旨在形成内部特征差异小、外部特征差异大的主题簇。

当前研究中所使用的聚类方法大多可以分为基于划分的方法、基于层次的方法、基于网格的方法、基于密度的方法、基于模型的方法 5 种类型。其中,基于划分的方法易于理解和实现且结果准确性高,但这类方法的准确性会受初始聚类数目的影响,例如 K-Means 聚类方法。基于层次的方法则不需要提前输入参数,但其计算复杂度较高且不能纠正错误的划分和合并,典型代表有 BIRCH 算法。基于网格的方法在运行过程中耗费的时间与数据量无关,但与每个维度上所划分的单元数相关,这也在一定程度上降低了聚类的质量和准确性。基于密度的方法可以发现任意形状的聚类,但运行中相关参数的设置对用户经验有一定要求。基于模型的方法可以自动修正划分中类的数目,但执行效率往往不高。

鉴于 LDA 模型提取结果数据量大且维度高的特点,本文最终选择基于划分的聚类方法。在计算主题间的相似性前,会利用特征向量对每一个主题进行表征,将主题之间距离的计算转换成对于特征向量之间距离的计算。因此,对于同为一种类型的文本内容,本文仅使用余弦相似度来测量主题之间的距离并进行后续的聚类分析。

3.2.3 主题演化持续性及其判定

在过去与主题演化相关的研究当中,学者通常以时间片段内与该主题相关的讨论存在与否作为其演化持续性的判定标准,即某一时间片段内与该主题相关的讨论数量不为零时可以认为该主题的演化仍然在持续。但在主题内容关联度的计算上普遍采用语义相似度或语义距离的方式,忽略了主题内容的时序特征及演化过程的有向性。除此以外,主题演化的相关研究在进行过程中往往选择将语料整体按照预定的设置分成若干个时间片段,却忽视了不同时间片段内语料本身的资源结构特征。因此,本文认为主题演化持续性的判定需要涉及主题强度、主题内容关联度和网络结构 3 个要素。其中,主题强度主要聚焦于从时序性上把握主题的变化趋势,主题内容关联度则更侧重于事件演化过程中主题内容在语义层面的深入或延伸,而网络结构贯穿于整个流程。已有的研究往往会选择将某一主题或事件相关的所有 UGC 数据视为单一整体,而忽略该整体内部的资源结构特征。但在实际情况中,内部的资源结构特征丰富了主题演化的方式,也为主题演化的探析提供了新的维度。

(1)主题强度。一方面,主题强度演化是指主题被关注的热点程度随时间的变化趋势,并以此刻画突发公共事件的生命周期[31]。另一方面,兼顾主题强度和主题内容两个层面的演化分析有助于提高对于主题演化持续性判断的准确性以及挖掘用户在事件演化期间的关注焦点。已有的与主题强度相关的研究文献也为本文提供了坚实的理论基础。本文将继续沿用已有的研究方法,在完成语料内容的主题分配后,对不同主题的语料数量进行统计,并以此作为主题强度的衡量依据并推进后续主题演化的分析。

(2)主题内容关联度。首先,内容关联度计算结果的准确性是影响该主题下讨论内容划分的关键因素。例如在数据分析过程中忽略了不同主题潜在的内容相关性,可能会出现内容相关的用户发言未被纳入分析范围、或关键词相同但所指代事件不同而被错误纳入分析范围从而影响主题演化分析结论的情况。其次,主题内容关联关系的指向也进一步完善了主题演化分析的内容。本文拟利用关联规则计算中所使用的支持度和可信度作为关系指向的判定标准,即实现主题演化分析的多维化。

(3)网络结构。作为反映语料内容资源结构的网络结构,是会话视角下主题演化分析需要纳入考虑的方面。会话视角下的研究强调对会话内容以及会话组织结构的研究,这也为主题演化分析提供了成熟的理论研究基础。随着关于主题演化以及有关 UGC 数据

应用方面的研究日趋成熟，越来越多的学者开始注意到网络结构本身的衍生性。主题演化的方式不再局限于参与讨论人数的变化，点赞、回复以及转发等用户表达意愿的行为方式也在影响着主题的演化。现如今大多数在线交流社区的交互模式逐渐呈现“主帖－回复帖－楼中楼帖”三级结构，包括知乎、新浪微博、百度贴吧等多个在线社区。所谓主帖，即用户首次在社区中发表与主题相关的言论。回复帖，则是用户自身或其他用户对于主帖的回复，此类回复帖大多是基于对于主帖内容的进一步衍生。楼中楼帖的形成原理与回复帖基本一致，即用户自身或其他用户对于回复帖内容的进一步补充。

4 突发公共事件持续演化过程分析

4.1 基于LDA的主题提取及主题簇生成

LDA模型主要通过词语共现概率来完成词语间的聚类，并利用狄利克雷分布对文档生成过程进行刻画。本文假定百度贴吧的帖子主题服从超参数狄利克雷先验分布，如公式(1)所示：

$$\mathrm{Dir}(\theta_c \mid \alpha) = \frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\prod_{t=1}^{T}\theta_{ct}^{\alpha_t-1} \qquad \text{公式(1)}$$

其中，θ_{ct}表示帖子c在主题t中的分布，对每一个生成的帖子主题t与主题词项之间服从分布$\varphi_t \sim Dir(\beta)$；对每篇帖子c与主题词之间服从分布$\theta_c \sim Dir(\alpha)$，对每篇帖子中的第n个词项生成主题项$z_{cn} \sim Multinomial(\theta_c)$和$w_{cn} \sim Multinomial(\varphi_{z_{cn}})$。基于此，本文的LDA似然模型如公式(2)所示：

$$p(W \mid \alpha,\beta) = \prod_{c=1}^{C}\int p(\theta_c \mid \alpha)\prod_{n=1}^{N_c}\sum_{z_c} p(z_{cn} \mid \theta_c)\, p(w_{cn} \mid \varphi_{z_{cn}})\, d\theta_c \qquad \text{公式(2)}$$

文本潜在主题数量设定的准确性也是影响主题模型提取准确性的关键因素，但LDA方法本身并不能自动生成最佳的主题数量。近年来的研究中不同的学者针对此问题提出了主题数量设定的不同方法或参考依据，例如困惑度(Perplexity)、非参数模型自动训练、Perplexity-var方法。但已提出的方法大多面临运算效率低、模型过拟合等问题。因此，本文借助一致性曲线确定最优主题数量，其计算过程大致可以划分为数据切分、概率计算、确认测度和取平均值4个步骤[32-33]。一致性分数在计算过程中通过融入布尔滑动窗口实现单词标记邻近性的捕获，旨在从语义层面分析文档内的特征词隶属于该主题的概率。

为了进一步降低LDA主题模型提取结果的稀疏性，本文选择基于余弦相似度的聚类算法作为主题簇生成实现的方法。该方法是一种基于相似度思想的聚类算法，测量相似度的指标是余弦值。作为基于划分的聚类方法，易于使用者理解和实现的同时，其聚类结果也呈现“高内聚低耦合”的特点。用余弦值来度量向量空间中两个向量之间的差异大小，更加关注到两个向量在方向上的差异，而不是距离或长度上的差异[34]。通过计算主题之间在内容维度上的余弦相似度来达到主题精确聚类的效果。具体计算公式见公式(3)：

$$cos(\theta) = \frac{\alpha * \beta}{|\alpha| * |\beta|} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \qquad \text{公式(3)}$$

其中，α和β均为n维向量。

首先利用提取出来的主题特征词构建一个词袋模型，并且针对词袋模型中的每一个词，参照主题的特征词，将主题表征为多个向量。然后依据构建得到的主题特征向量对主题两两之间进行余弦相似度计算，得到主题之间的相似度值。根据相似度计算结果，为不同类型帖子选取最合适的主题簇数量。

以主帖层级的帖子为例，最终得到的聚类结果见图2。在LDA主题模型提取完成后得到的结果中仍然存在区分度不高的情况。以主题感染表现为例，其与主题体征表现尽管在模型抽取结果中为两个独立主题，但与包括新闻报道在内的其他主题相比，这两者的主题内容存在一定程度的相似性。诸如此类的情况将会导致主题过于分散的现象，从而会进一步影响后续主题强度计算的准确性。

4.2 基于一致性分数的主题分配

在得到若干个主题簇及其相对应的主题后，仍然需要界定帖子的主题归属，即每一个帖子属于哪一个或哪些主题。本文在主题分配的过程中参照的标准是每条帖子中每个主题的一致性分数，即每个主题与该帖子内容层面的匹配程度。若在分配过程中出现两个主题隶属于该帖的概率相同的情况，则通过人工判断来赋予主题。主题分配完成后，结合主题簇生成结果，为每一个帖子分配对应的主题簇，以实现更精确的主题分配。

考虑到语料内容的资源结构，在帖子主题分配的过程中，不同层级结构的帖子对应该层级结构的主题。即当为主帖分配主题的时候，需要从主帖中提取出的候选主题中进行分配，与其他两个层级的帖子主题无关。界定完成后，建立主题－主题簇的对应关系，并统计不同主题簇下的帖子数量作为后续基于主题强度分析的数据来源。

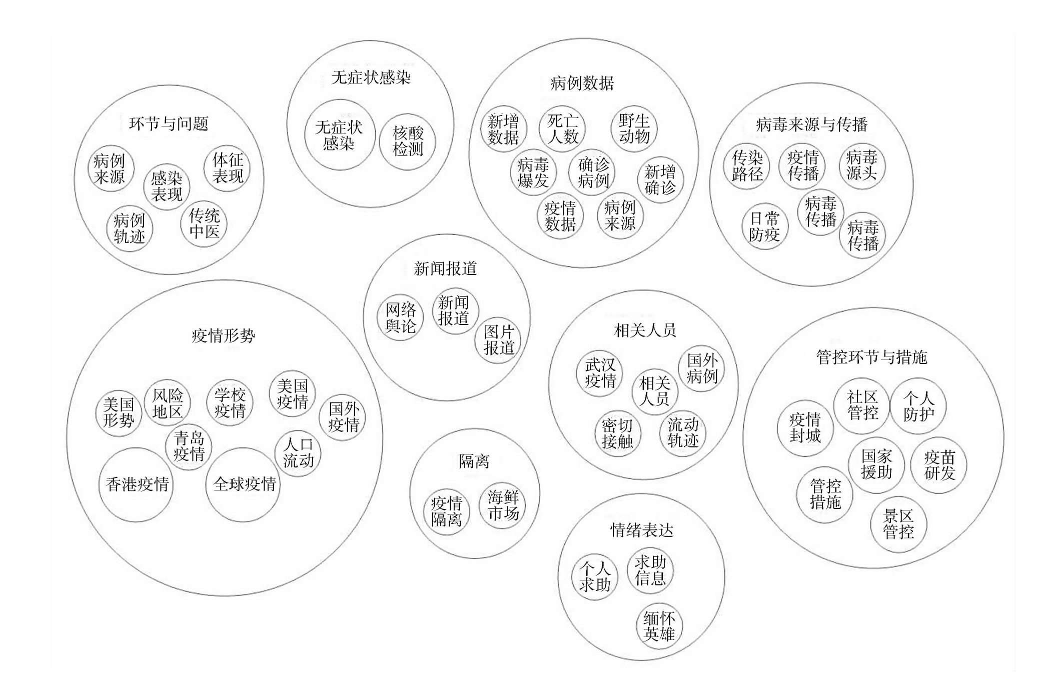


图 2　基于主题聚类的主帖层级主题簇生成结果

4.3　基于时序性与主题强度的突发公共事件主题演化

基于时序性与主题强度的突发公共事件主题演化主要在时序维度上分别进行主题簇热度和主题簇间交互性的分析。

关于主题簇热度的分析，本文通过对每个主题簇下不同主题在不同时间切片内出现的频次进行求和来反映该主题簇在不同时段内的讨论热度。通过针对主题簇热度演化趋势的分析，总结突发公共事件主题演化的规律。

前文所提到的主题分配策略的主要思想是为每一个帖子分配其对应一致性分数最高的主题。这一思想所建立的前提条件是每一个帖子仅属于某一主题或主题簇。但在实际情况中存在帖子对应多个主题的情况。因此，为进一步研究主题簇之间的交互关系，本文将每个帖子分配的主题数量设定由一个增加至三个，以期望通过研究主题簇之间的共现关系来充分体现交互关系，即若主题簇之间的共现次数越高，则可以说明两者存在更为紧密的交互性。

4.4　基于关联规则的突发公共事件主题演化

关联分析作为实现知识发现的常见手段，一般用于量化描述物品 A 的出现在多大程度上依赖于物品 B 的出现[35]。将关联分析应用于突发公共事件主题演化中，旨在有效揭示特征词与特征词之间的依存关系，如某一关于药物治疗的文本中，包含“药品 B”这一特征词的同时，也存在“疾病 A”特征词，则表明“药品 B”可能具有治疗“疾病 A”的功效。并据此可进一步获得由若干特征词所构成的主题与主题间的语义关联关系，从而达到分析该主题演化路径的目的。因此，根据研究目的，本文将每一个帖子视作一个事务 T，其由多个特征词的项目组成，为获取特征词之间的语义关联，可对其构建共现矩阵从而实现其关联分析，在满足一定支持度和可信度条件下挖掘出频繁出现在一起的特征词。

完成主题提取后，假设存在 T 个帖子，每个帖子中包含 N 个互相独立的特征词，且这 N 个特征词均需要从该帖所分配的主题中获取，即若该帖中出现特征词并不属于该帖所分配主题的特征词集合时，则不纳入计算范围。

设帖子数为 T，首先计算每个特征词出现的帖子数量 $C(C \leqslant N)$，C_A 表示含有特征词 A 的主题数；再计算任意两个特征词共现的帖子数，记为 R；最后计算支持度 S、可信度 Co。其中，支持度表示特征词 A 和特征词 B 共同出现在所有帖子中的概率，计算方法见公式(4)。可信度表示特征词 A 出现的帖子中，特征词 B 出现的概率，计算方法见公式(5)：

$$S_{(A \to B)} = \frac{R}{T} \qquad \text{公式(4)}$$

$$Co_{(A \to B)} = P(B \mid A) = \frac{R}{C_A} \qquad \text{公式(5)}$$

由于支持度只能说明特征词 A、B 同时出现的概率，并不能量化特征词 A、B 之间的关联关系强度，故本

文将支持度作为判别条件，识别大于等于最小支持度的有向词对{$A\to B$}，并在此基础上，只对已识别的强关联词对进行可信度计算，据此作为主题 – 主题关联分析的依据。本文进一步对特征词 – 特征词之间的关联类型进行界定，将其分为 3 种类型，即前序关系、后继关系以及平行关系，如表 1 所示：

表 1　特征词之间的关联关系类型

关系	向量表示	解读
前序关系	只存在 $A\to B$，可信度为 $Co_{(A\to B)}$	特征词 A 的出现影响特征词 B 的出现，定义 A 为 B 的前序特征词，B 为 A 的后继特征词
后继关系	既存在 $A\to B$ 关系又存在 $B\to A$ 的关系，且 $Co_{(A\to B)} > Co_{(B\to A)}$	特征词 A 对特征词 B 出现的影响大于特征词 B 对特征词 A 出现的影响，舍弃 $B\to A$，定义 A 为 B 的前序特征词，B 为 A 的后继特征词
平行关系	既存在 $A\to B$ 关系又存在 $B\to A$ 的关系，且 $Co_{(A\to B)} = Co_{(B\to A)}$	特征词 A 与特征词 B 之间的影响相同，具有双向关系，可视作等同关系进行考量

由于特征词是用来表征主题的向量，本文通过计算不同主题下的各特征词间的支持度总和并取平均值的方式来衡量主题间的关联关系强度。即假设存在包含特征词 A、B、C 的主题 1 和包含特征词 D、E、F 的主题 2。则主题 1 和主题 2 的关联关系强度 $Co_{(tp1\to tp2)}$ 和 $Co_{(tp2\to tp1)}$ 计算方法见公式(6)和公式(7)：

$$Co_{(tp1\to tp2)} = (Co_{(A\to D)} + Co_{(A\to E)} + Co_{(A\to F)} + Co_{(B\to D)} + Co_{(B\to E)} + Co_{(B\to F)} + Co_{(C\to D)} + Co_{(C\to E)} + Co_{(C\to F)} \div 9$$

公式(6)

$$Co_{(tp2\to tp1)} = (Co_{(D\to A)} + Co_{(D\to B)} + Co_{(D\to C)} + Co_{(E\to A)} + Co_{(E\to B)} + Co_{(E\to C)} + Co_{(F\to A)} + Co_{(F\to B)} + Co_{(F\to C)} \div 9$$

公式(7)

关联强度矩阵的构建有助于对主题进行进一步的演化分析。完成主题间的关联强度计算后，本文将主题间的关联强度转化为一模矩阵。首先，同一主题间不存在演化的情况时将同一主题之间的可信度设置为 0。其次，对于主题间的可信度 $Co_{(tp1\to tp2)}$ 和 $Co_{(tp2\to tp1)}$，会预先对比二者的大小，较小的可信度用 0 替代。例如当 $Co_{(tp1\to tp2)}$ 大于 $Co_{(tp2\to tp1)}$ 时，证明主题 2 的出现受到了主题 1 的影响，那么此时主题 1 对于主题 2 的影响可以忽略不计。若出现 $Co_{(tp1\to tp2)}$ 等于 $Co_{(tp2\to tp1)}$ 的情况，证明主题 1 和主题 2 是等价关系，保留双方原有可信度。本文所构建的矩阵包括同层级帖中的主题 – 主题关联矩阵构建和跨层级帖中的主题 – 主题关联矩阵构建两种类型，旨在将网络结构进一步纳入演化分析的范畴。

5　实证分析

5.1　数据获取与预处理

本文选择百度贴吧的“新型冠状病毒”吧作为采集对象，采集数据的时间段为该吧成立后一年内，即 2020 年 1 月 21 日至 2020 年 12 月 21 日，共采集 52 025 条数据，其中包括主帖 7 298 条、回复帖 20 049 条、楼中楼帖 24 678 条。为了保证实验数据集的完整性、清洁性和结构化，本文通过繁简体转换、删除空值以及纯字符串数据以实现前期语料处理，然后调用 Python 中的 jieba 分词数据包，结合哈工大停用词表对发帖内容进行进一步的分词处理，将处理后的帖子内容分别保存在不同字段中，成为后续用于候选主题提取的有效数据。

5.2　主题及主题簇生成

数据采集和清洗完成后，最终得到有效数据 39 563 条，其中包括主帖 7 280 条、回复帖 17 950 条、楼中楼帖 14 333 条。在主题模型训练过程中，通过绘制一致性曲线最终确定输出主题数为 50，并输出主题以及对应特征词及其在该主题语义内容表达层面的贡献概率。以主帖中的“疫情传播”主题为例，其具体输出结果如表 2 所示：

表 2　“疫情传播”主题提取输出结果

主题	特征词及概率
疫情传播	0.305 * “疫情” + 0.075 * “结束” + 0.068 * “严重” + 0.053 * “这次” + 0.032 * “可能” + 0.016 * “发展” + 0.014 * “非典” + 0.011 * “原因” + 0.010 * “告诉” + 0.007 * “特殊” + 0.007 * “两年” + 0.006 * “年前” + 0.005 * “妈妈” + 0.005 * “外包装” + 0.004 * “病床” + 0.004 * “万人” + 0.004 * “排名” + 0.004 * “后人” + 0.004 * “再次” + 0.004 * “或超”

根据构建得到的主题特征向量对主题两两之间进行余弦相似度计算，依据相似度计算结果，为每一个帖子选取最合适的主题数量。在主题及主题簇的生成过程中，本文通过已有的特征词及概率分布结果，对主题进行人工归纳并命名，最终生成基于不同层级的主要主题簇以及对应主题如表 3 所示。以情绪表达这一主题簇为例，尽管在主帖、回复帖和楼中楼帖 3 个层级的语料内容中均涉及该主题簇，但在主帖层级中，与该主题簇相关的用户发言更侧重于对于自身需求的表达和对于抗疫前线英雄的缅怀；在回复帖层级中，该主题簇下的发言内容则更聚焦于对于疫情好转态势的美好向

往；而在楼中楼帖层级中，围绕该主题簇的用户发言更倾向于对国家或地区防疫政策的赞美。

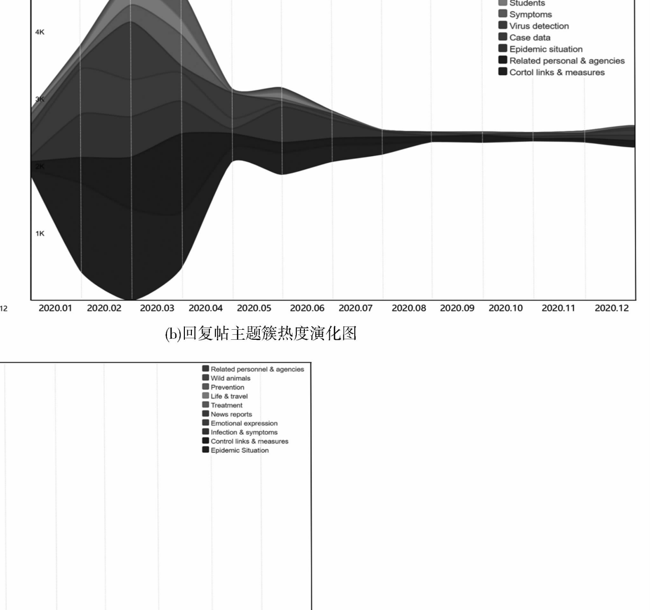
表 3　基于不同层级的主要主题簇及对应主题

主题簇	主帖主题	回复帖主题	楼中楼帖主题
环节与问题	病例来源（T1）/感染表现（T2）/体征表现（T5）/病例轨迹（T12）/传统中医（T31）		
病毒来源与传播	传染路径（T3）/疫情传播（T7）/病毒源头（T18）/病毒传播（T22）/日常防疫（T23）/病毒传播（T41）		
病例数据	病例来源（T0）/死亡人数（T8）/新增数据（T10）/病毒暴发（T16）/确诊病例（T27）/新增确诊（T38）/疫情数据（T48）/野生动物（T49）	患者治疗（P2）/疫情数据（P10）/无症状感染（P12）/病例诊断（P18）/确诊病例（P19）/全球疫情（P25）/具体病例（P29）/确诊病例（P36）/临床确诊（P37）/病例筛查（P46）/医院隔离（P48）	
情绪表达	求助信息（T13）/个人求助（T26）/缅怀英雄（T28）	祈福行为（P4）/情绪表达（P14）	情绪表达（C1）/情绪表达（C8）/赞美国家（C20）/诽谤造谣（C28）/负面造谣（C40）
疫情形势	美国形势（T29）/青岛疫情（T30）/学校疫情（T32）/美国疫情（T33）/国外疫情（T36）/人口流动（T37）/风险地区（T39）/全球疫情（T43）/香港疫情（T44）	美国疫情（P3）/负面影响（P11）/黑龙江疫情（P15）/日本疫情（P16）/疫情数据（P10）/美国政策（P26）/美国病例（P27）/国外形势（P34）/美国形势（P38）	疫情传播（C4）/武汉形势（C14）/武汉现状（C16）/美国疫情（C24）/国外形势（C25）/日本（C26）/疫情源头（C37）/传播方式（C38）/病例信息（C43）
管控环节与措施	疫情封城（T17）/国家援助（T19）/疫苗研发（T20）/个人防护（T24）/管控措施（T25）/社区管控（T35）/景区管控（T47）	防控措施（P1）/国外防控（P6）/防控工作（P21）/消毒工作（P23）/个人防护（P30）	管控环节（C0）/核酸检测（C3）/社区管控（C5）/防疫要求（C10）/国家管控（C33）/隔离措施（C36）/口罩佩戴（C42）
病毒发现		瓷器传播（P0）/仓库（P9）/市场商户（P40）/野生动物（P24）	
生活出行			出行问题（C2）/学生开学（C17）/生活问题（C27）/物流恢复（C49）

5.3　基于时序性与主题强度的演化分析

根据所收集到的数据结果，本文以月为单位进行时间片段的划分，以主题簇的讨论热度为纵轴，以时序为横轴，分别绘制三个层级帖子主题簇的热度演化趋势图以及对应的交互演化图，如图 3 和图 4 所示：

(a)主帖主题簇热度演化图



(b)回复帖主题簇热度演化图

(c)楼中楼帖主题簇热度演化图

图 3　各层级主题簇热度演化情况

(a)主帖主题簇间交互演化图

(b)回复帖主题簇间交互演化图

(c)楼中楼帖主题簇间交互演化图

图4　各层级主题簇间交互演化情况

本文选择河流图来展示不同主题簇的演化及交互情况，其中不同主题簇河流的宽度代表对应时间点的主题簇热度，即主题簇河流所占纵轴比例越大，其讨论热度越高。两条河流交叉即代表主题簇之间存在共现关系，即具有交互性。

从主帖这一层级来看，在主题簇热度演化方面，2020年4月和6月附近10个主题簇均出现了两个峰值。从单个时间片段来看，疫情形势和管控环节与措施这两个主题簇分别占据了4月和6月两个时间片段内的最大讨论热度。2020年4月8日，武汉市正式解封的消息成为了4月这一峰值产生的契机；而导致6月这一峰值产生的原因则是中国新冠疫苗启动三期临床试验这一信息，三期临床试验的结果直接影响到疫苗研发的成功与否。在交互分析方面，主帖主题簇间的交互演化也呈现“先增强后减弱”的特点。从交叉次数来看，演化初期主题簇间交互程度低，差异度大。随着事件讨论热度的上升和下降，不同主题簇的共现频率出现了明显的提升或减弱。

从回复帖这一层级来看，在主题簇热度演化方面，回复帖主题簇的演化趋势与主帖层级有所不同，其演化趋势呈现“先增强后减弱”的特点，整体上并没有出现二次峰值点。而在主题簇间交互性的演化方面，回复帖主题簇间的交互性相对较差，随着主题在演化后期的消亡导致交互性逐渐减弱甚至于不存在。从楼中楼帖这一级来看，楼中楼帖的主题簇热度呈现整体下降的趋势。从交互性来看，尽管演化前期交叉频率高，不同主题簇间的交互性强，但后期受到讨论热度下降的影响，主题簇间的交互性逐渐减弱。

通过对比3个层级帖子主题簇热度的演化规律，本文认为影响主题簇讨论热度和交互性的演化因素包括标志性事件的出现、资源结构关系和事件特性3个方面。

首先是标志性事件，主帖主题簇演化热度出现两个峰值的原因均是标志性新闻的出现打破了原有的事件演化趋势，再次吸引用户重新投入讨论中。其次是资源结构关系，排除标志性事件的影响，3个层级帖子主题簇热度演化图谱趋势呈现明显的滞后性。这其实是受到了百度贴吧自身的“主帖－回复帖－楼中楼帖”层级结构的影响。随着层级结构的深入，越来越多的用户会参与到讨论之中，这就导致了处于层级结构最深处的楼中楼帖在自身的演化初期就达到了讨论热度峰值情况的出现。最后是事件本身的特性，与一般公共事件所不同，突发性公共事件的特性之一就是突发性，即毫无征兆的突然发生。主题簇热度演化图谱

中的骤增或骤减趋势均充分反映了突发公共事件的这一特性。

5.4 基于关联规则的主题及主题簇演化分析

5.4.1 主帖－回复帖－楼中楼帖主题演化规律分析

依据4.4中提出的关于关联规则的计算方法，本文共计得到1 448个特征词的支持度。单个特征词的支持度是衡量一个词在整个主题文档集合中重要性的指标，即某词的支持度越高，该词在文档集合中越重要。特征词对的支持度用于衡量两个词共同出现在所有主题文档集合的概率，支持度越高，则说明两个词关联度越高且该词对也具有不可被忽视的重要意义。除支持度以外，可信度用于界定特征词之间的关联关系类型，并用于衡量主题之间的关联关系。

在实际分析过程中，尽管前期提取了一定数量的具有平行关系的特征词对，但在主题关联关系分析结果中却主要呈现为前序或后继的演化关系。本文利用NetDraw软件绘制不同层级主题内部及外部的演化关系并得到可视化图谱，如图5所示：

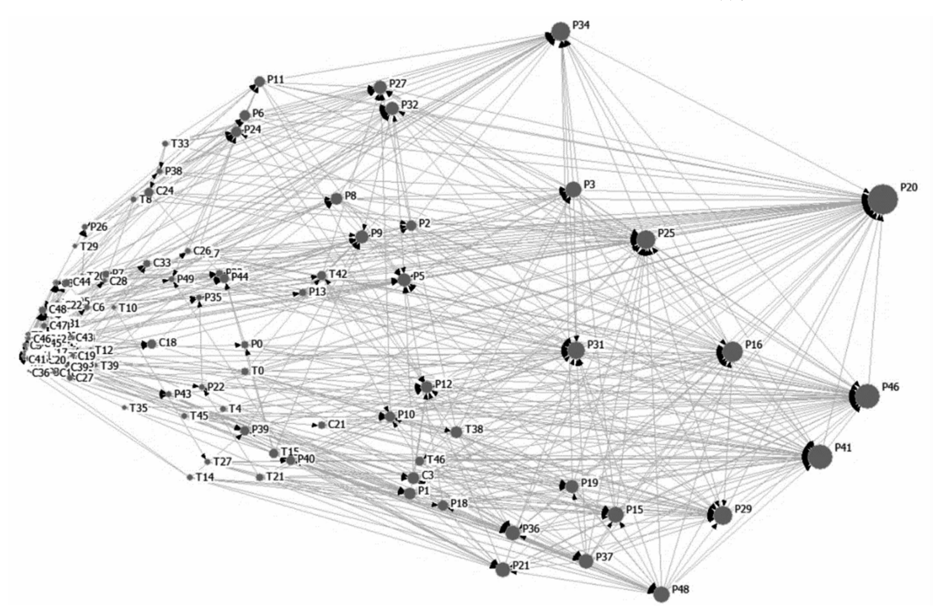


图5 3个层级帖子主题间演化关系

其中，T、P、C分别表示主题所属层级，即主帖、回复帖和楼中楼帖。例如T5用于表示从主帖中提取出的编号为5的主题。箭头的指向则用于表示主题之间的演化关系。例如主题关系“T5 = > C30”即表示楼中楼帖主题30的出现是参照了主帖主题5。图中圆圈面积的大小即代表该主题的度中心性，即有色圆圈面积越大，代表该主题在演化过程中重要程度越高。

从整体来看，回复帖的度中心性普遍较高，P46、P20、P25具有一定的代表性。对于主帖而言，T38的度中心性相对较高，且大多指向其他层级的节点。对于楼中楼帖而言，C3的度中心性相对较高，且处于大多数连线的箭头终点。这与资源结构本身的特性有关。在“主帖－回复帖－楼中楼帖”三层结构中，主帖成为了新的主题产生点，回复帖承担了中间的过渡和发散作用，而楼中楼帖则承担了总结和深化前两层级主题内容的作用。但鉴于LDA所提取出的主题密度较为稀疏，为了进一步挖掘关键主题演化路径，本文对前文构建的主题簇结果进行演化分析。

5.4.2 主帖－回复帖－楼中楼帖主题簇演化规律分析

参照4.4中根据特征词对的关联关系强度计算主题间关联关系强度的方式，将各主题视为用于表示对应主题簇的特征词，并基于此将主题间的关联关系强度进一步映射到主题簇层面来研究主题簇间的演化关系，利用Neo4j工具得到主题簇演化关系，见图6。

在关系图谱中，本部分继续沿用前文设定的字母T、P、C用于表示主题簇所属层级，即主帖、回复帖和楼中楼帖，而主题簇间的关联关系主要通过连线来体现。通过分析，本文最终确定新型冠状病毒肺炎疫情相关帖子中的核心主题簇为主帖层级的“相关人员”“疫情形势”“病例数据”，回复帖层级的“症状”“疫情形式”“病例数据”，楼中楼帖层级的“管控环节与措施”“感染及症状”。根据核心主题簇以及对应主题簇下的主题演化关系，最终确定3条关键演化路径，分别是主帖层级的“相关人员”至回复帖层级的“疫情形势”至楼中楼帖层级的“管控环节与措施”、主帖层级的“疫情

翟姗姗，王左戎，陈欢，等.会话分析视角下的突发公共事件主题演化研究——以“新冠肺炎疫情”为例[J].图书情报工作,2022,66(11):87－99.

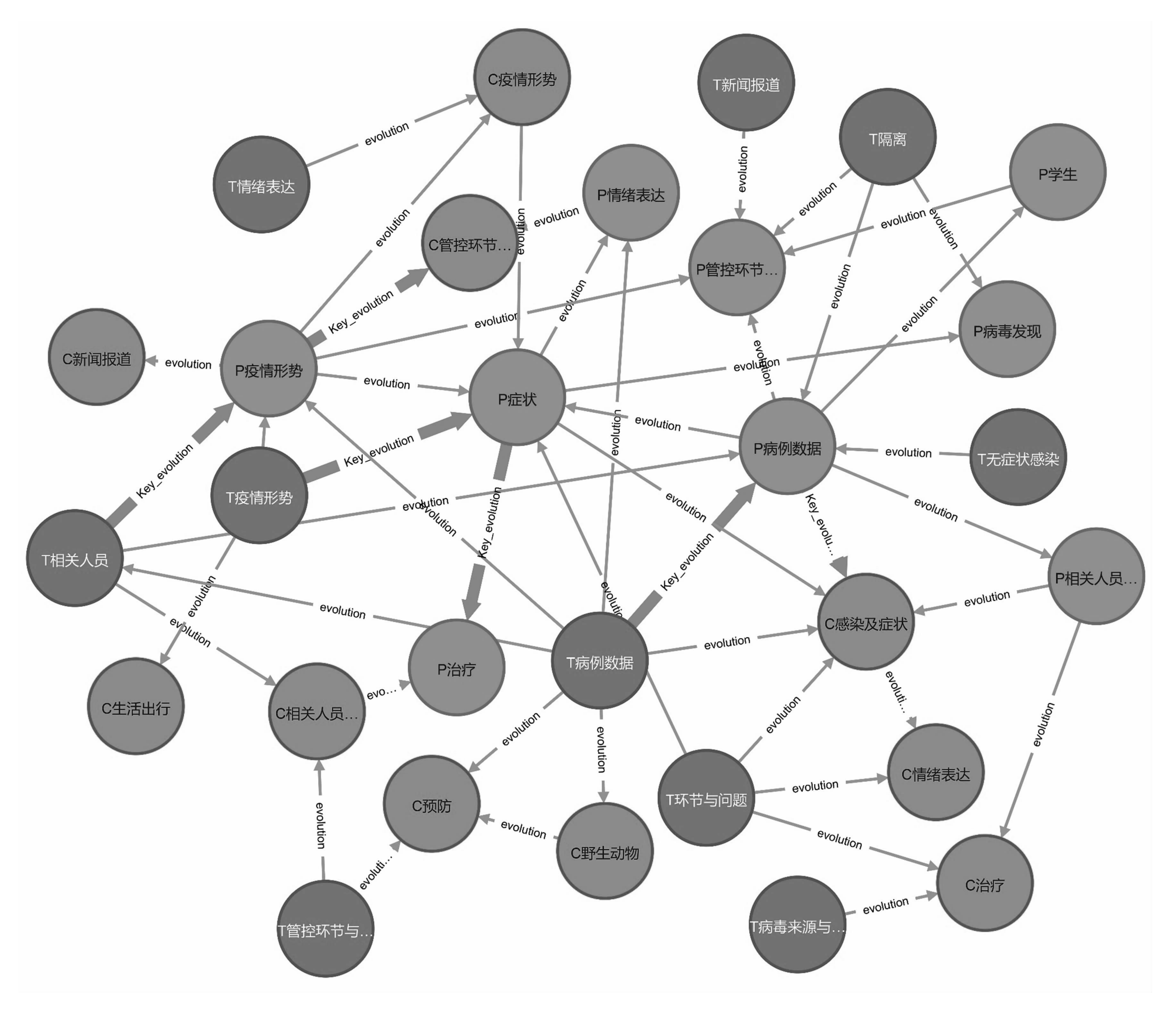


图6　3个层级核心主题簇演化关系图谱

形势”至回复帖层级的“症状”至同样是回复帖层级的“治疗”、主帖层级的“病例数据”至回复帖层级的“病例数据”至楼中楼帖层级的“感染及症状”。

除此以外，本文将会话视角下主题簇间的演化路径归纳为6种，如表4所示。结合关键演化路径分析的结果，本文认为资源结构也是影响主题簇间演化的关键因素。过去关于主题演化的研究中，学者们在采集数据时只关注直接发布的内容，即本文中涉及的“主帖”内容，或者将3个层级的内容视为一个帖子的全部内容进行后续分析。但根据本文从主题或主题簇层面的演化分析结果来说，处于回复帖层级的帖子内容度中心性普遍较高，成为了影响主题演化方向的重要角色。这也为相关部门提供了一个全新的舆情控制方向，即关注帖子中的评论内容，而绝不仅关注那些影响力较高的用户发言。

6　结语

本文在主题演化相关的研究基础上融入了会话分析的视角，提出主题演化持续性的判定应从主题强度、网络结构和主题内容关联度3个方面来进行。在演化分析中，一方面兼顾主题强度和时序性，通过描绘并对比不同层级帖子主题簇热度和交互性变化，提出影响主题强度和交互性演化的三大因素，分别为标志性事件的出现、网络结构和事件特性。另一方面基于本文提出的关联规则，利用特征词对的支持度和可信度来对主题簇间的关系进行进一步判断，同时应用社会网络分析方法抓取核心主题簇，挖掘与其相关的关键演化路径，为相关部门舆情疏导提供参考性方向。

表4　主题簇间演化关系及具体实例

演化模式	演化路径	具体实例(主题簇－主题簇)
层级内演化	主帖－主帖	病例数据－相关人员
	回复帖－回复帖	病例数据－疫情形势
	楼中楼帖－楼中楼帖	感染及症状－管控环节与措施
层级间演化	主帖－回复帖	疫情形势－症状
	主帖－楼中楼帖	相关人员－相关人员与机构
	回复帖－楼中楼帖	疫情形势－管控环节与措施

参考文献：

[1] 张宁熙. 大数据在突发公共事件网络舆情信息工作中的应用[J]. 现代情报, 2015, 35(6): 38 - 42.

[2] 马超,翟姗姗,王晓. 会话分析视角下非正式信息交流主题与主题簇演化分析[J]. 图书情报工作, 2021, 65(17): 91 - 100.

[3] RITZHAUPT A D, STEWART M, SMITH P, et al. An investigation of distance education in North American research literature using co-word analysis[J]. International review of research in open and distance learning, 2010, 11(1): 37 - 60.

[4] 巴志超,李纲,朱世伟. 共现分析中的关键词选择与语义度量方法研究[J]. 情报学报, 2016, 35(2): 197 - 207.

[5] 王红斌,王健雄,张亚飞,等. 主题不平衡新闻文本数据集的主题识别方法研究[J]. 数据分析与知识发现, 2021, 5(3): 109 - 120.

[6] 李真,丁晟春,王楠. 网络舆情观点主题识别研究[J]. 数据分析与知识发现, 2017, 1(8): 18 - 30.

[7] 邱宁佳,杨长庚,王鹏,任涛. 改进卷积神经网络的文本主题识别算法研究[J]. 计算机工程与应用,2022,58(2):161 - 168.

[8] 邵琦,牟冬梅,王萍,等. 基于语义的突发公共卫生事件网络舆情主题发现研究[J]. 数据分析与知识发现, 2020, 4(9): 68 - 80.

[9] 安璐,杜廷尧,李纲,等. 突发公共卫生事件利益相关者在社交媒体中的关注点及演化模式[J]. 情报学报, 2018, 37(4): 394 - 405.

[10] 李月. 突发公共卫生事件中公共政策主题演化研究——以国家中心城市官方微信为例[J]. 情报杂志, 2020, 39(9): 143 - 149.

[11] 曹树金,岳文玉. 突发公共卫生事件微博舆情主题挖掘与演化分析[J]. 信息资源管理学报, 2020, 10(6): 28 - 37.

[12] 赵蓉英,常茹茹,陈湛,等. 基于知乎平台的突发公共卫生事件主题演化研究[J]. 信息资源管理学报, 2021, 11(2): 52 - 59.

[13] 姚艾昕,马捷,林英,等. 重大突发公共卫生事件谣言演化与治理策略研究[J]. 情报科学, 2020, 38(7): 22 - 29.

[14] 刘雅姝,张海涛,徐海玲,等. 多维特征融合的网络舆情突发事件演化话题图谱研究[J]. 情报学报, 2019, 38(8): 798 - 806.

[15] 徐月梅,吕思凝,蔡连侨,等. 结合卷积神经网络和 Topic2Vec 的新闻主题演变分析[J]. 数据分析与知识发现, 2018, 2(9): 31 - 41.

[16] 朱晓霞,宋嘉欣,孟建芳. 基于动态主题——情感演化模型的网络舆情信息分析[J]. 情报科学, 2019, 37(7): 72 - 78.

[17] 张光陆. 深度学习视角下的课堂话语互动特征:基于会话分析[J]. 中国教育学刊, 2021(1): 79 - 84.

[18] 王亚峰,于国栋. 医患交流中患者扩展回答的会话分析研究[J]. 外语教学理论与实践, 2021, 175(3): 108 - 118.

[19] 胡文芝,廖美珍. 中国心理治疗话语"解述"现象的会话分析研究[J]. 重庆大学学报(社会科学版), 2013, 19(4): 92 - 100.

[20] 沈芮妃. 谈话节目《圆桌派》的会话结构[J]. 青年记者, 2017, 566(18): 72 - 73.

[21] 张黎. 现场促销员的会话策略分析[J]. 语言文字应用, 2007, 63(3): 87 - 93.

[22] 巴志超,李纲,毛进,等. 微信群内部信息交流的网络结构、行为及其演化分析——基于会话分析视角[J]. 情报学报, 2018, 37(10): 1009 - 1021.

[23] 李纲,李显鑫,巴志超,等. 微信群会话网络结构及成员角色划分研究[J]. 现代情报, 2018, 38(7): 3 - 11.

[24] 卢恒,张向先,张莉曼,等. 会话分析视角下虚拟学术社区用户交互行为特征研究[J]. 图书情报工作, 2020, 64(13): 80 - 89.

[25] 李月琳,张建伟,张婳. 螺旋式与直线式:在线健康医疗平台用户与医生交互模式研究[J]. 情报学报, 2021, 40(1): 88 - 100.

[26] 桂小庆,张俊,张晓民,等. 时态主题模型方法及应用研究综述[J]. 计算机科学, 2017, 44(2): 46 - 55.

[27] BLEI D M, LAFFERTY J D. Dynamic topic models[C]//Proceedings of the 23rd international conference on machine learning. New York: ACM, 2006: 113 - 120.

[28] WANG X R, MCCALLUM A. Topics over time: a non-markov continuous-time model of topical trends[C]//Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2006:424 - 433.

[29] CHENG X, YAN X, LAN Y, et al. BTM: topic modeling over short texts[J]. IEEE transactions on knowledge and data engineering, 2014, 26(12): 2928 - 2941.

[30] ZUO Y, ZHAO J, XU K. Word network topic model: a simple but general solution for short and imbalanced texts[J]. Knowledge and information systems, 2016, 48(2): 379 - 398.

[31] 关鹏,王曰芬,傅柱. 基于 LDA 的主题语义演化分析方法研究——以锂离子电池领域为例[J]. 数据分析与知识发现, 2019, 3(7): 61 - 72.

[32] SYED S, SPRUIT M. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation[C]//2017 IEEE international conference on data science and advanced analytics (DSAA). Tokyo: IEEE, 2017: 165 - 174.

[33] RÖDER M, BOTH A, HINNEBURG A. Exploring the space of topic coherence measures[C]//Proceedings of the eighth ACM international conference on Web search and data mining. New York: ACM, 2015:399 - 408.

[34] 成怡,朱伟康,徐国伟. 基于余弦相似度的改进 ORB 匹配算法[J]. 天津工业大学学报, 2021, 40(1): 60 - 66.

[35] 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002.

作者贡献说明：

翟姗姗:确定选题;提出论文框架,论文修改;
王左戎:撰写论文;
陈欢:处理数据;
潘港辉:获取及预处理数据。

翟姗姗，王左戎，陈欢，等. 会话分析视角下的突发公共事件主题演化研究——以“新冠肺炎疫情”为例[J]. 图书情报工作,2022,66(11):87－99.

Research on Topic evolution of Public Emergencies from the Perspective of Conversation Analysis: A Case Study of COVID-19

Zhai Shanshan　Wang Zuorong　Chen Huan　Pan Ganghui

School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [Purpose/Significance] The introduction of conversation analysis theory provides a new research perspective for the study of topic evolution and refines the granularity of topic evolution analysis. At the same time, a more perfect theme evolution analysis approach is applied to public emergencies, which is conducive to improving the efficiency of public opinion guidance of regulatory departments. [Method/Process] Based on the topic identification methods and topic evolution judgment criteria in existing studies, this paper combined conversation analysis and topic analysis to introduce conversation contents and conversation organization structure into the process of topic evolution analysis, and used user-generated content in COVID-19 as data source for empirical analysis. Through the topic evolution analysis based on temporal sequence and discussion hot, the evolution laws of contents at different levels were identified from the topic intensity level. The association rule calculation idea of knowledge discovery was introduced at the topic content analysis level, to mine the reference relationship between corpus contents, and the key evolution path was determined by combining the social network analysis method. [Result/Conclusion] The results show that there are certain differences in the topic contents at different levels in the network structure and it has an important influence on the evolution trend of the topic, effective supervision of the contents at important levels will play a positive role in guiding the trend of public opinion.

Keywords: conversational analysis　public emergency　topic identification　topic evolution　association rules

刘国钧先生相关资料征集启事

刘国钧先生(1899－1980),字衡如,其一生阅历丰富,交游广泛,治学勤奋,取得了丰硕的学术成果,是中国近现代图书馆学的奠基人之一和著名学者。为编纂《刘国钧全集》,全面反映刘国钧先生为中国图书馆事业做出的杰出贡献,《刘国钧全集》编纂课题组现面向社会各界公开征集刘国钧先生相关资料及线索。

一、征集范围及内容

包括且不限于以下类型:

1、手稿、书信及题签等相关手迹;

2、记录刘国钧先生工作、生活和活动的照片、音像资料;

3、各个时期的著述、出版物;

4、印章;

5、其他与刘国钧先生相关的史料(例如藏书等)。

6、其他与刘国钧先生相关的一切资料线索。

二、征集载体

1、各类文字记录、电子文件、照片、声像、字画和实物等。

2、如不能提供原件,烦请以电子邮件形式寄送扫描件(PDF 或 JPEG 图片格式,内容清晰)

三、征集时间

长期征集,时间不限,随时联系。

四、资料利用

课题组对所征集的资料、实物等将妥善保存、合理利用;对提供的线索,课题组将派人专门联系、整理。征集资料收入《刘国钧全集》者,将在书中注明资料来源,以示谢忱。

五、联系方式

联 系 人:张久珍

通信地址:北京海淀区颐和园路 5 号北京大学信息管理系(方李邦琴楼 406 室)

邮　　编:100871

电　　邮:jiu@ pku. edu. cn

电　　话:010－62766306

《刘国钧全集》编纂项目管理组

2022 年 4 月 7 日